Appendix

Theorem 1 (FC-bound) *For any set $\mathbb{X}$, any $\epsilon \in [0,1]$, and a fixed classifier a such that $m = n(a,\mathbb{X})$ the probability of overfitting is given by the left tail of the hypergeometric distribution:*

$$Q_\epsilon(a,\mathbb{X}) = H_L^{\ell,\,m}\left(\tfrac{\ell}{L}(m-\epsilon k)\right). \tag{1}$$

Proof: Denote $s = n(a,X)$ and rewrite the overfitting condition $\delta(a,X) \geq \epsilon$ as $\frac{1}{k}(m-s) - \frac{1}{\ell}s \geq \epsilon$ or equivalently $s \leq \frac{\ell}{L}(m-\epsilon k) \equiv s_m(\epsilon)$. Then

$$Q_\epsilon = \mathsf{P}\big[n(a,X) \leq s_m(\epsilon)\big] = \sum_{s=s_0}^{\lfloor s_m(\epsilon)\rfloor} \mathsf{P}\big[n(a,X)=s\big] = \sum_{s=s_0}^{\lfloor s_m(\epsilon)\rfloor} h_L^{\ell,\,m}(s) = H_L^{\ell,\,m}(s_m(\epsilon)).$$

■

Theorem 2 (VC-bound) *For any set $\mathbb{X}$, any learning algorithm μ, and any $\epsilon \in [0,1]$ the probability of large uniform deviation is bounded by the sum of FC-bounds over the set A:*

$$\widetilde{Q}_\epsilon(A,\mathbb{X}) \leq \sum_{a\in A} H_L^{\ell,\,m}\left(\tfrac{\ell}{L}(m-\epsilon k)\right), \quad m = n(a,\mathbb{X}). \tag{2}$$

Proof: Apply a union bound substituting the maximum of binary values by their sum:

$$\widetilde{Q}_\epsilon = \mathsf{P}\max_{a\in A}\big[\delta(a,X)\geq\epsilon\big] \leq \sum_{a\in A}\mathsf{P}\big[\delta(a,X)\geq\epsilon\big] = \sum_{a\in A} H_L^{\ell,\,m}(s_m(\epsilon)), \quad m = n(a).$$

■

Further weakening gives a well known form of the VC-bound:

$$\widetilde{Q}_\epsilon(A,\mathbb{X}) \leq |A| \max_m H_L^{\ell,\,m}(s_m(\epsilon)) \leq |A|\cdot\tfrac{3}{2}e^{-\epsilon^2\ell}, \quad \text{if } \ell = k,$$

where $|A|$ is called a *shattering coefficient* of the set of classifiers A on the set $\mathbb{X}$.

The principle of protective and prohibitive subsets

The principle of protective and prohibitive sets [3] is based on the conjecture that the necessary and sufficient condition for $\mu X = a$ can be specified explicitly for any classifier $a \in A$ in terms of subsets of objects. From this conjecture an exact Q_ϵ bound has been derived.

In this work we use a similar conjecture relaxed to the necessary condition and derive an upper bound which has a simpler form.

Theorem 3 *For each classifier $a \in A$ there exists a* protective subset $X_a \in \mathbb{X}$ *and a* prohibitive subset $X'_a \in \mathbb{X}$ *such that for any $X \in [\mathbb{X}]^\ell$*

$$\big[\mu X = a\big] \leq \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]. \tag{3}$$

The subset $\mathbb{X}\backslash X_a \backslash X'_a$ is called *neutral* for a classifier a. The presence or absence of neutral objects in a training sample X does not change the result of learning μX. Later we will give nontrivial examples of μ and A that satisfy conjecture 3.

Lemma 4 *If conjecture 3 holds, then the probability to learn a classifier a can be bounded:*

$$\mathsf{P}\big[\mu X = a\big] \leq P_a \equiv \tbinom{L_a}{\ell_a}/\tbinom{L}{\ell},$$

where $L_a = L - |X_a| - |X'_a|$ and $\ell_a = \ell - |X_a|$ are the number of neutral objects for a classifier a in the general set $\mathbb{X}$ and sample X respectively.

Proof: According to the conjecture $\mathsf{P}\big[\mu X = a\big] \leq \mathsf{P}\big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]$. The right-hand side is a fraction of partitions $\mathbb{X} = X \sqcup \bar{X}$ such that $X_a \subseteq X$ and $X'_a \subseteq \bar{X}$. The number of such partitions is equal to $\binom{L_a}{\ell_a}$. The number of all partitions is equal to $\binom{L}{\ell}$, hence their ratio gives P_a. ■

Theorem 5 *If conjecture 3 holds, then for any $\epsilon \in [0,1]$ the bound on probability of overfitting is*

$$Q_\epsilon \leq \sum_{a \in A} P_a H_{L_a}^{\ell_a,\, m_a}\left(s_a(\epsilon)\right), \tag{4}$$

where $m_a = n(a, \mathbb{X} \backslash X_a \backslash X'_a)$ is a number of errors that classifier a produces on neutral objects and $s_a(\epsilon) = \frac{\ell}{L}\big(n(a) - \epsilon k\big) - n(a, X_a)$ is a largest number of errors $n(a, X \backslash X_a)$ that classifier a produces on neutral training objects provided that discrepancy $\delta(a, X)$ exceeds ϵ.

Proof: The probability of overfitting Q_ϵ can be found as a total probability from probability to learn each of classifiers $\mathsf{P}\big[\mu X = a\big]$ and conditional probabilities $Q_{\epsilon|a} = \mathsf{P}\big[\delta(a, X) \geq \epsilon \mid a = \mu X\big]$:

$$Q_\epsilon = \sum_{a \in A} \mathsf{P}\big[\mu X = a\big] Q_{\epsilon|a} \leq \sum_{a \in A} P_a Q_{\epsilon|a}.$$

The conditional probability $Q_{\epsilon|a}$ can be obtained from theorem 2 by taking into account that the subsets X_a and X'_a can not be involved in partitioning given a fixed classifier a. Only L_a neutral objects are partitioned into ℓ_a training and $L_a - \ell_a$ testing objects. To employ theorem 2 we express the discrepancy $\delta(a, X)$ in terms or the number of errors on neutral training objects $s = n(a, X \backslash X_a)$:

$$\delta(a, X) = \tfrac{1}{k}\big(n(a) - s - n(a, X_a)\big) - \tfrac{1}{\ell}\big(s + n(a, X_a)\big).$$

Condition $\delta(a, X) \geq \epsilon$ is equivalent to $s \leq s_a(\epsilon)$. Then $Q_{\epsilon|a} = H_{L_a}^{\ell_a,\, m_a}\left(s_a(\epsilon)\right)$ and (4) holds. ■

Note that the sum $\sum_a P_a$ can be interpreted as a degree of looseness of the bound (4). The bound is exact if this sum is equal to 1.

The splitting and connectivity graph

Define an order relation on classifiers $a \leq b$ as a natural order over their error vectors: $a_i \leq b_i$ for all $i = 1, \ldots, L$. Define a metric on classifiers as a Hamming distance between error vectors: $\rho(a, b) = \sum_{i=1}^{L} |a_i - b_i|$. Classifiers a and b are called *connected* if $\rho(a, b) = 1$. Define the precedence relation on classifiers $a \prec b$ as $\big(a \leq b\big) \wedge \big(\rho(a, b) = 1\big)$.

The set of classifiers A can be represented by a multipartite directed graph $\langle A, E \rangle$ that we call the *splitting and connectivity graph* (SC-graph) in which vertices are classifiers, and edges (a, b) are pairs of classifiers such that $a \prec b$, see example on Figure 1. The partite subsets $A_m = \{a \in A \colon n(a) = m\}$ are called *error layers*, $m = 0, \ldots, L$. Each edge of the SC-graph (a, b) corresponds to an object $x_{ab} \in \mathbb{X}$ such that $I(a, x_{ab}) = 0$ and $I(b, x_{ab}) = 1$.

SC-graph is much the same as 1-inclusion graph used in [1] to obtain lower bounds on VC-dimension. The VC-dimension may result in highly overestimated generalization bounds as it is based on the union bound. In our combinatorial framework the SC-graph is used to replace the union bound by a much more accurate technique.

Note that SC-graph can be considered also as a subgraph of the Hasse diagram (the graph of transitive reduction) of the partial order over error vectors.

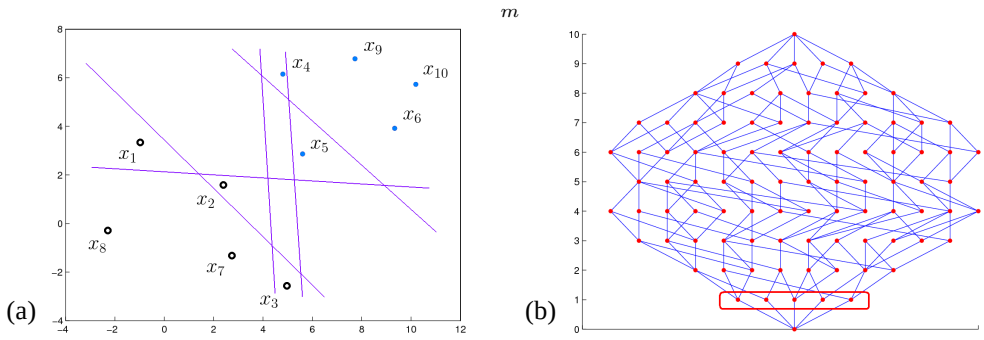
SC-bound for pessimistic Empirical Risk Minimization

Lemma 6 *If learning algorithm μ is pessimistic ERM, then conjecture 3 holds, and for any $a \in A$*

$$X_a = \big\{x_{ab} \in \mathbb{X} \;\big|\; a \prec b\big\} \;\; \textit{is the protective subset};$$
$$X'_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A \colon b \leq a,\; I(b, x) < I(a, x)\big\} \;\; \textit{is the prohibitive subset}.$$

Proof: Let us use a proof by contradiction showing that if $\mu X = a$, then $X_a \subseteq X$ and $X'_a \subseteq \bar{X}$.

Assume that an object $x_{ab} \in X_a$ not belonging to X exists. Then $n(a, X) = n(b, X)$ because the error vectors a and b differ by exactly one object x_{ab}. At the same time $n(a, \mathbb{X}) + 1 = n(b, \mathbb{X})$, therefore the learning algorithm μ being pessimistic learns the classifier b rather than a from the training sample X which contradicts the initial condition $\mu X = a$. Then we conclude that $X_a \subseteq X$.



(c)

	layer 0	layer 1					layer 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	…
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	…
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	…
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	…
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	…
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	…
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	…
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…

Figure 1: Two-dimensional linearly separable classification task with $L = 10$ objects of 2 classes and 5 linear classifiers that produce exactly one error (a). The SC-graph over the set of all 2-dimensional linear classifiers (b). The first layer ($m = 1$) corresponds to 5 classifiers shown at the left chart. The fragment of error matrix corresponding to layers $m = 0, 1, 2$ (c).

Assume that an object $x \in X'_a$ belonging to X exists. Then $n(b, X) < n(a, X)$. The learning algorithm μ being empirical risk minimizer learns the classifier b rather than a from the training sample X which contradicts the initial condition $\mu X = a$. Then we conclude that $X'_a \subseteq \bar{X}$. ■

Corollary 7 *Any classifier $a \in A$ produces errors on all prohibitive objects X'_a and does not produce errors on all protective objects X_a.*

Upper connectivity $q(a) = |X_a|$ of a classifier a is the *out-degree* of the vertex a in the SC-graph, i. e. the number of edges leaving the vertex a.

Lower connectivity $d(a) = |X'_a|$ of a classifier a is the *in-degree* of the vertex a in the SC-graph, i. e. the number of edges entering the vertex a.

Inferiority $r(a) = |X'_a|$ of a classifier a is the number of different objects assigned to edges below the vertex a in the SC-graph. If a correct classifier $a_0 \in A$ exists such that $n(a_0) = 0$, then inferiority is equal to the number of errors, $r(a) = n(a)$. In general case, $d(a) \leq r(a) \leq n(a)$.

Theorem 8 (SC-bound) *If learning algorithm μ is ERM, then for any $\epsilon \in [0, 1]$ the probability of overfitting is bounded by the weighted sum of FC-bounds over the set A:*

$$Q_\epsilon(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{\binom{L-q-r}{\ell-q}}{\binom{L}{\ell}} H^{\ell-q,\, m-r}_{L-q-r}\left(\tfrac{\ell}{L}(m - \epsilon k)\right), \qquad (5)$$

where $q = q(a)$ is upper connectivity, $r = r(a)$ is inferiority, $m = n(a)$ is the number of errors of classifier a on the general object set $\mathbb{X}$.

Proof: The bound (5) for pessimistic ERM follows immediately from theorem 5, lemma 6, and corollary 7. Then (5) also holds for any ERM. ■

The weight $P_a = \binom{L-q-r}{\ell-q} / \binom{L}{\ell}$ in the sum (5) is an upper bound on the probability to learn the classifier a. Its value decreases exponentially as connectivity $q(a)$ and inferiority $r(a)$ increase. This fact has two important consequences.

First, connected sets of classifiers are less subjected to overfitting. Note that an attempt to use only the fact of connectedness with no counting the number of connections did not lead to a tight bound [2].

Second, only a little part of lower layers contribute significantly to the probability of overfitting. This fact encourages effective procedures for level-wise bottom-up SC-bound computation.

The SC-bound (5) is much more tight than the VC-bound (2). It can be transformed into the VC-bound by substituting $q = r = 0$, i. e. by totally disregarding the SC-graph structure.

References

[1] Haussler D., Littlestone N., Warmuth M. K. *Predicting $\{0, 1\}$-functions on randomly drawn points* // *Inf. Comput.* — December 1994. — Vol. 115. — Pp. 248–292.

[2] Sill J. *Monotonicity and connectedness in learning systems*: Ph.D. thesis / California Institute of Technology. — 1998.

[3] Vorontsov K. V. *Exact combinatorial bounds on the probability of overfitting for empirical risk minimization* // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 269–285.